Epidemiologic Perspectives & Innovations

BioMed Central

Methodology

Open Access

Covariate balance in a Bayesian propensity score analysis of beta blocker therapy in heart failure patients

Lawrence C McCandless*[1], Paul Gustafson[2], Peter C Austin[3,4,5] and Adrian R Levy[6]

Address: [1]Faculty of Health Sciences, Simon Fraser University, Canada, [2]Department of Statistics, University of British Columbia, Canada, [3]Institute for Clinical Evaluative Sciences, Toronto, Canada, [4]Dalla Lana School of Public Health, University of Toronto, Canada, [5]Department of Health Policy, Management and Evaluation, University of Toronto, Canada and [6]School of Population and Public Health, University of British Columbia, Canada

Email: Lawrence C McCandless* - mccandless@sfu.ca; Paul Gustafson - gustaf@stat.ubc.ca; Peter C Austin - peter.austin@ices.on.ca; Adrian R Levy - adrian.levy@ubc.ca

* Corresponding author

Published: 10 September 2009

Epidemiologic Perspectives & Innovations 2009, **6**:5 doi:10.1186/1742-5573-6-5

Received: 11 December 2008
Accepted: 10 September 2009

This article is available from: http://www.epi-perspectives.com/content/6/1/5

Abstract

Regression adjustment for the propensity score is a statistical method that reduces confounding from measured variables in observational data. A Bayesian propensity score analysis extends this idea by using simultaneous estimation of the propensity scores and the treatment effect. In this article, we conduct an empirical investigation of the performance of Bayesian propensity scores in the context of an observational study of the effectiveness of beta-blocker therapy in heart failure patients. We study the balancing properties of the estimated propensity scores. Traditional Frequentist propensity scores focus attention on balancing covariates that are strongly associated with treatment. In contrast, we demonstrate that Bayesian propensity scores can be used to balance the association between covariates and the outcome. This balancing property has the effect of reducing confounding bias because it reduces the degree to which covariates are outcome risk factors.

Introduction

Regression adjustment for the propensity score is a statistical method that reduces confounding from measured variables in observational data. The idea is to use the propensity score, defined as the probability of treatment given measured confounders, to build treatments groups that are similar with respect to outcome risk factors [1]. Patients with the same propensity score have the same distribution of measured confounders. Provided that there is no unmeasured confounding, we obtain unbiased estimates of the treatment effect by comparing treatment groups within levels of the propensity score. Analytic techniques using propensity scores include stratifying on quintiles of the propensity score, or using the propensity score as a covariate in a regression model for the outcome. Other methods using matching or weighting are available [2].

Recently, McCandless, Gustafson and Austin [3] proposed a statistical method which combines regression adjustment for the propensity score with Bayesian techniques. Their proposed Bayesian propensity score analysis (BPSA) models the propensity score as a latent variable that is integrated from the posterior distribution for the treatment effect. BPSA fits regression models for the outcome and treatment simultaneously rather than one at a time.

When estimating the propensity scores, BPSA incorporates prior information about the relationship between the outcome and propensity score within treatment groups. In contrast, standard analytic methods estimate propensity scores from the marginal model for treatment given measured confounders.

Other Bayesian techniques using propensity scores are given by Hill and McCulloch [4] and Hoshino [5]. The approach of Hill and McCulloch [4] uses nonparametric modelling of the outcome using Bayesian additive regression trees. It has the advantage that the user is not required to supply modeling assumptions about the manner in which variables are parametrically related. Hoshino describes a Markov chain Monte Carlo technique for fitting propensity models to observational data [5].

McCandless et al. [3] use simulations to evaluate the performance of BPSA. They demonstrate that if the regression model for the relationship between outcome and propensity score is correctly specified, then BPSA permits more efficient estimation of the propensity scores compared to other non-Bayesian methods. However, when the outcome regression model is incorrectly specified, this can adversely impact BPSA and give propensity score estimates that are asymptotically biased. Thus it is unclear whether BPSA will outperform standard non-Bayesian approaches in real data applications. In practice, statistical models for the outcome variable are only approximations.

An appealing feature of propensity score techniques is that simple diagnostics tools have been developed to compare the performance of competing propensity score estimates for control of confounding. Conditional on the propensity score, treatment and confounders are independent. The distribution of the confounders should be similar across treatment groups. This can be empirically verified by comparing summary statistics such as the mean and variance for covariates in treatment versus control. If the distributions are similar then this indicates that treatment effect estimates are unconfounded. Following convention in the literature, we refer to this diagnostic procedure as checking covariate *balance*. A detailed discussion is given by Austin and Mamdani [2].

The balancing properties of propensity score estimates have been well studied. Austin and Mamdani [2] and Austin et al. [6] studied the impact of different variable selection strategies on the balancing properties of estimated propensity scores. The authors show that including non-confounders in the model for the propensity score can reduce the amount of covariate balance on the confounders. Matching on the propensity score produces greater balance compared to stratifying on quintiles of the propensity score. Austin et al. [7] investigated covariate balance in settings where there are unmeasured confounders.

The logic of checking covariate balance provides an opportunity to evaluate the performance of BPSA. We can empirically verify if the propensity score estimates yield similar balance compared to conventional methods. Accordingly, our objective is to study covariate balance for BPSA. In what follows, we present a case-study of an observational study of the effectiveness of beta-blocker therapy in British Columbia heart failure patients. In the example, there is strong confounding because beta-blockers are preferentially prescribed to younger, healthier patients. We analyze the data using BPSA and compute propensity estimates from the posterior distribution of the propensity scores. We study balance with respect to treatment when stratifying on the estimated propensity scores. Our analysis reveals that BPSA gives worse balance compared to conventional propensity scores estimates calculated from the marginal model for treatment. The covariate distributions differ in treatment versus control. However, we then show that BPSA yields improved in balance with respect to the outcome. By this we mean that stratifying on BPSA propensity score estimates reduces the strength of the association between the covariates and mortality. This reduces confounding bias because it reduces the impact of covariate imbalances between treatment groups. Thus BPSA makes a tradeoff between balancing baseline covariates with respect to the treatment versus the outcome variable. In contrast, conventional propensity score estimates are calculated from the marginal model for treatment. They handle all confounders equally, regardless of whether they are important outcome risk factors.

Estimating the effectiveness of beta-blocker therapy in heart failure patients

To investigate the ability of BPSA to control confounding, we consider the example of an observational study of the effect of beta-blockers on one year all-cause mortality in heart failure patients from British Columbia. Beta-blockers are a class of cardiovascular therapies which act on the beta-adrenergic nervous system to improve heart function [8]. Randomized trials show that they reduce in mortality in heart failure patients, but there is interest in quantifying the magnitude of this effect within the general population, including the very elderly [9]. In Canada, beta-blockers are more often prescribed to patients who are young, healthy and with fewer comorbidities [10]. Because treated patients in the population are healthier than untreated patients, we expect that a crude comparison of mortality rates will be confounded and tend to exaggerate the benefits of beta-blocker therapy. Treated patients will have lower mortality because they are in better health, even in the absence of any benefit of beta-blockers.

In this study, we obtained administrative health data for one year of follow-up on 6969 patients discharged from British Columbia hospitals in 1999 and 2000. Using records of hospitalization and drug prescription claims, we compiled information on demographic characteristics, comorbid medical conditions and medications dispensed from community pharmacies throughout the province. Vital status at the end of follow up was established by electronic linkage of medical records to death certificates. Full details are provided elsewhere [8,11]. After one year, 1755 patients died, and the mortality rate among treated patients was 19% versus 27% among untreated patients. The crude odds ratio for the association between beta-blockers and mortality is 0.64 with 95% credible interval (0.55, 0.75). In contrast, meta-analyses of randomized controlled trials consistently report a 30% reduction in mortality with beta blocker use [9]. This suggests that the association between beta-blocker therapy and mortality is confounded due to measured and unmeasured indications for disease severity. To estimate the treatment effect analytic adjustments are required, and this provides a test case for comparing the performance of BPSA with other methods.

Bayesian propensity score analysis: Data, models and estimation

Let X denote a binary variable representing exposure to beta-blockers. We set X equal to one if the subject was dispensed a beta-blocker within 30 days of discharge from hospital and zero otherwise. The binary response variable Y is set equal to one if the subject died within one year of discharge from hospital and zero otherwise. Let $C = (C_1, C_2,..., C_p)$ denote a vector of $p = 21$ potential confounding variables measured on or before hospital discharge including demographic characteristics: age (categorical with four levels; <65, 65-74, 75-84, ≥ 85 years), sex (binary with one indicating female and zero otherwise); indicator variables for comorbid conditions: cerebrovascular disease, chronic obstructive pulmonary disorder (COPD), hyponatremia, metastatic disorder, renal disease, ventricular arrhythmia, liver disease, malignancy, shock; indicator variables for dispensation of heart failure medications within thirty days of hospital discharge: angiosin converting enzyme (ACE) inhibitors, angiotensin II receptor blockers (ARB), calcium channel blockers (CCB), digoxin, diuretics, statins; and characteristics of the index hospitalization: indicator of transferred status, hospital length of stay in days. In order to ease the specification of intercept terms in regression modelling, we set the first component of C (denoted C_0) to be equal to one.

To model the propensity score and relationship between Y, X and C, we use two logistic regression models. Following McCandless et al. [3], we let

$$\text{logit}\{Pr(Y = 1 \mid X, C)\} = \xi_0 + \sum_{j=1}^{l=3} \xi_j g_j\{z(C, \gamma)\} + \beta X \quad (1)$$

$$\text{logit}\{Pr(X = 1 \mid C)\} = \gamma^T C. \quad (2)$$

The quantity β models the treatment effect, while the parameter $\gamma = (\gamma_0,..., \gamma_p)$ is a $(p + 1) \times 1$ vector of regression coefficients which identifies the propensity score, given by $Z = \text{logit}\{Pr(X = 1|C)\} = \gamma^T C$. Following Rubin and Thomas [12], we define the propensity score as the log odds of treatment given measured confounders. This definition differs slightly from the usual definition in the literature, but it eases the analytical tractability of studying propensity score estimates. In practice, both definitions give similar treatment effect estimates because the log odds transformation is monotonic [12].

In equation (1), we use regression splines to flexibly model the nonparametric relationship between the propensity score and outcome variable. In the summation $\sum_{j=1}^{l=3} \xi_j g_j\{z(C, \gamma)\}$, the quantities $g_j\{\cdot\}$, $j = 1,..., l = 3$ are natural cubic spline basis functions with $l = 3$ knots (q_1, q_2, q_3), and regression coefficients $\xi = (\xi_1, \xi_2, \xi_3)$. The choice of $l = 3$ knots reflects a trade off between smoothness and complexity. Alternatively, we could use hierarchical models to model uncertainty in the location or number of knots.

We assign prior distributions for the model parameters β, γ, ξ as

$$\beta \sim N(0, \sigma_\beta^2)$$
$$\gamma_0, \ldots, \gamma_p \sim N(0, \sigma_\gamma^2)$$
$$\xi_0, \ldots, \xi_3 \sim N(0, \sigma_\xi^2),$$

where $\sigma_\beta^2 = \sigma_\gamma^2 = \sigma_\xi^2 = \{\log(15)/2\}^2$. The value for σ_β^2 models the belief that the odds ratio for the treatment effect is not overly large and lies between 1/15 and 15 with probability 95%. The values for σ_γ^2 and σ_ξ^2 make similar modelling assumptions about the prior magnitude for the association between Y and Z given X, and also the association between C and X.

The regression models in equations (1) and (2) give a likelihood function for the data. Combining the likelihood and prior distributions, we can sample from the posterior

distribution of the treatment effect β and nuisance parameters γ, ξ using Markov chain Monte Carlo. Conceptionally, the implementation involves a two step iterative procedure: First, impute the propensity score parameter γ. Second, fit a complete data step to estimate the treatment effect β and ξ given the propensity scores. Successive iterations average over uncertainty in the propensity scores. The approach has close connections to the EM algorithm and multiple imputation. Computer code for implementing BPSA in the software package R [13] is available [see additional file 1]. A detailed discussion of implementing BPSA is given in McCandless et al. [3].

Before applying BPSA to the data, we first select the knots used in the spline regression for the relationship between mortality and propensity score. To choose the knots, we fit the logistic regression model given in equation (2) via maximum likelihood and compute the fitted values. The three knots are chosen as $q_1 = 0.10$, $q_2 = 0.18$, $q_3 = 0.24$, which define quartiles of the estimated propensity scores. We then apply BPSA to the data by sampling from the posterior density $P(\beta, \xi, \gamma|data)$. We run a single MCMC chain of length 100 000 after discarding 10 000 initial iterations. Sampler convergence is assessed by simulating separate MCMC chains with overdispersed starting values and the diagnostic tools supplied in the CODA package in R [13].

Analysis results

The results are given in Table 1 under the heading "BPSA", which contain posterior means and 95% credible intervals for the treatment effect β and the regression coefficients γ. We omit estimates of ξ because the quantity is a nuisance parameter with an interpretation that depends on the parameterization of the natural splines in equation (1).

While the priors distributions are plausible, they may nonetheless be informative. We repeat the analysis by fixing the prior variances equal to 10^3 rather than $\{\log(15)/2\}^2$. Additionally, we experiment with uniform priors bounded on the interval [-10, 10]. The resulting inferences for β are similar to those in Table 1 and differ by less than 0.03 on the log odds scale. For the parameter γ, the MCMC output is similar under different priors, although there is less shrinkage towards origin using the uninformative priors. Posterior means differed by at most 0.05 for all covariates except metastatic disorder and malignancy.

Table 1: Log odds ratios (95% CIs) for the treatment effect β and the regression coefficients γ calculated using BPSA and PSA.

Description	Parameter	Log Odds Ratio (95% Interval Estimate)	
		BPSA	PSA
Beta blocker	β	-0.21 (-0.37, -0.05)	-0.31 (-0.46, -0.15)
Demographics			
Female Sex	γ_1	0.17 (0.06, 0.29)	0.12 (-0.01, 0.25)
Age			
< 65 (reference)	.	0.00	0.00
65 - 74	γ_2	-0.19 (-0.32, -0.04)	-0.09 (-0.3, 0.12)
75 - 84	γ_3	-0.40 (-0.56, -0.24)	-0.21 (-0.41, 0.00)
> 85	γ_4	-0.71 (-0.94, -0.46)	-0.37 (-0.59, -0.14)
Comorbid conditions			
Cerebrovascular dis.	γ_5	-0.11 (-0.67, 0.44)	0.25 (-0.46, 0.96)
COPD	γ_6	-0.32 (-0.60, -0.06)	-0.89 (-1.30, -0.49)
Hyponatremia	γ_7	-0.02 (-0.26, 0.21)	0.03 (-0.33, 0.39)
Metastatic disorder	γ_8	-1.42 (-2.33, -0.56)	-0.40 (-1.37, 0.57)
Renal disease	γ_9	-0.17 (-0.32, 0.01)	0.38 (0.15, 0.62)
Ventricular arrhythmia	γ_{10}	-0.12 (-0.63, 0.44)	0.12 (-0.62, 0.86)
Liver disease	γ_{11}	-0.52 (-1.03, -0.08)	-1.11 (-2.04, -0.19)
Malignancy	γ_{12}	-0.78 (-1.19, -0.34)	-0.06 (-0.57, 0.45)
Shock	γ_{13}	-0.06 (-0.56, 0.39)	-0.12 (-0.83, 0.58)
Hospitalization			
Transferred	γ_{14}	-0.41 (-0.58, -0.25)	-0.01 (-0.22, 0.20)
Stay (10 day intvs.)	γ_{15}	-0.13 (-0.18, -0.09)	-0.05 (-0.11, 0.01)
Heart failure medications			
Digoxin	γ_{16}	-0.02 (-0.11, 0.07)	0.00 (-0.14, 0.13)
Diuretic	γ_{17}	0.28 (0.09, 0.48)	0.72 (0.54, 0.90)
CCB	γ_{18}	0.22 (0.08, 0.35)	0.27 (0.10, 0.44)
ACE inhibitor	γ_{19}	0.29 (0.11, 0.45)	0.61 (0.47, 0.76)
ARB	γ_{20}	0.18 (-0.06, 0.46)	0.53 (0.19, 0.87)
Statin	γ_{21}	0.91 (0.65, 1.24)	0.94 (0.76, 1.12)

For comparison, we also apply a propensity score analysis (PSA) to the data. We define PSA as the following two step procedure: First, fit the logistic regression model in equation (2) by maximum likelihood and compute the estimated propensity scores from the fitted values. Next, fit the model in equation (1) by maximum likelihood, substituting the fitted values in place for the true propensity scores. PSA is a standard method for controlling confounding [2]. It is identical to BPSA, except that it fits the regressions models in equations (1) and (2) one at at time rather than simultaneously. PSA is implemented using the same knots (q_1, q_2, q_3) as BPSA. The results are given in the second column of Table 1 under the heading "PSA".

As expected, the estimates for the treatment effect β from BPSA and PSA are less than zero and the interval estimates exclude zero, indicating that beta-blockers reduce mortality in heart failure patients. But the treatment effect estimates are slightly different. BPSA gives an odds ratio of exp(-0.21) = 0.81 whereas PSA gives exp(-0.31) = 0.73. BPSA and PSA also give different inferences for γ. The differences in point estimates of γ are substantial, although they are generally small compared to the 95% interval estimates. From equation (2), we can see that the parameter γ models the propensity score in the sense that if γ is known then the propensity score for a patient with covariate vector C is given by $\gamma^T C$. Because the estimates of γ differ for BPSA versus PSA, this suggests that the propensity score estimates also differ.

Which inferences should we prefer? BPSA and PSA use identical models, but yield qualitatively different answers. The reason is because BPSA fits regression models for Y and X simultaneously rather than one at at time. McCandless et al. [3] conducted detailed simulations and showed that if the outcome model in equation (1) is sufficiently non-parametric to capture the dependence between Y and Z, then BPSA gives more efficient estimates of the propensity scores. But for the heart failure data the true data generating process is unknown.

To explore the results in greater detail, we compare the estimated propensity scores from either method. Let $\hat{\gamma}^{PSA}$ denote the estimate for γ obtained from PSA and let $\hat{\gamma}^{BPSA}$ denote the posterior mean for γ from BPSA. The estimated propensity score from PSA is $\hat{Z}^{(PSA)} = \hat{\gamma}^{(PSA)T} C$, whereas for BPSA it is $\hat{Z}^{(BPSA)} = \hat{\gamma}^{(BPSA)T} C$. Figure 1 plots $\hat{Z}^{(BPSA)}$ versus $\hat{Z}^{(PSA)}$ for a random sample of 1000 subjects in the study. The quantities have correlation equal to 0.85, but their dependence is nonetheless heterogeneous. The linear clustering in the figure is due to the covariate for hospital length of stay. This is the only continuous covariate in the dataset, with median length of stay equal to 5 days and interquartile range of 3 to 10 days. We see in Table 1 that BPSA and PSA give different estimates for γ_{15} which models the relationship between length of stay and treatment assignment. The clusters in Figure 1 are groups of patients who spent different amounts of time in hospital, but otherwise have the same covariate pattern.

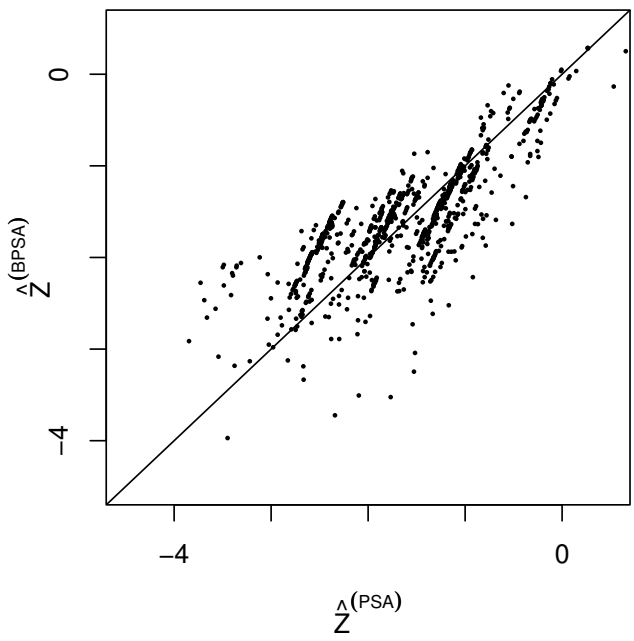

Figure 1
$\hat{Z}^{(BPSA)}$ versus $\hat{Z}^{(PSA)}$ for a random sample of 1000 subjects in the heart failure study.

In PSA we control confounding by stratifying on $\hat{Z}^{(PSA)}$, whereas in BPSA we stratify on $\hat{Z}^{(BPSA)}$. Figure 1 shows that the methods stratify subjects into different groups. Provided that the models in equations (1) and (2) are correct, then large sample Bayesian theory tells us that the parameter estimates for γ calculated from BPSA and PSA will be asymptotically identical and consistent to the true parameter value [14]. But for the heart failure data, the combination of a finite sample size and possible model misspecification leads to sizeable differences in the propensity score estimates.

One of the attractive properties of propensity score techniques is that simple diagnostic tools are available to study the performance of competing propensity score estimates. If we condition on the propensity score, then the treatment and confounders are independent. The empiri-

cal distribution of the confounders should be balanced across treatment groups, and this can be verified by comparing the mean and variance of covariates in treatment versus control. If the distributions are similar then this indicates that the confounding has been reduced [12]. We use the notion of covariate balance as a starting point for evaluating BPSA versus PSA.

Balance with respect to treatment

In this section, we investigate the balancing properties of $\hat{Z}^{(BPSA)}$ and $\hat{Z}^{(PSA)}$. Rosenbaum and Rubin [1] showed that $X \perp C|Z$, where the symbol "$\perp$" means that X and C are conditionally independent given the true propensity score Z. Stratifying on the propensity score confers *balance with respect to treatment* and breaks the association between X and C. To investigate the performance of competing propensity score estimates, we adopt an approach similar to Imai and Van Dyk [15] and fit models of the form

$$\text{logit}\{P(C_k = 1 \mid X, Z)\} = \phi_k + \sum_{j=1}^{l=3} \omega_{jk} g_j\{Z\} + \theta_k X \quad \text{for } k = 1, \ldots, 21 \tag{3}$$

where C_k denotes the k^{th} component of C. Equation (3) is identical to equation (1) except that it substitutes each of the covariates in place of the outcome variable Y. To understand the logic behind fitting such a model to assess balance, notice that if $\hat{Z}$ is equal to the true propensity score, then this implies that $\theta_1 = \theta_2 = \theta_3 = \ldots = \theta_{21} = 0$ in equation (3) because $X \perp C_k | \hat{Z}$ for each of $k = 1,\ldots, 21$. Thus the extent to which estimates of $\theta_1,\ldots, \theta_{21}$ depart from zero speaks to the balancing properties of competing propensity score estimates, and therefore, the effectiveness of BPSA and PSA in control of confounding. If $X \perp C_k | \hat{Z}$, then 95% interval estimates for θ_k should cover zero with probability 95%.

This reasoning is analogous to investigating balance by reporting covariate summary statistics within quintiles of the estimated propensity scores. The original approach of assessing balance recommended by Rosenbaum and Rubin [1] proceeds as follows: First, estimate the propensity score via regression of treatment on covariates. Next, break the population into five separate quintile groups based the estimated propensity scores. Lastly, check balance within each quintile group by comparing the distribution of covariates (e.g. age) in treated versus untreated. If the distributions are similar, then this suggests that the estimated propensity scores succeed in breaking the association between treatment and confounders. Equation (3) assesses balance in a similar fashion. We regress C_k on X and $\hat{Z}$. If the regression coefficient θ_k is zero, then this means that X and C_k are not associated after having stratified on $\hat{Z}$. Therefore $\hat{Z}$ is a "good" propensity score estimate because it induces balance with respect to treatment.

To illustrate in the heart failure data, we begin by studying the crude associations between X and C. This is accomplished by fitting the 21 regressions in equation (3) while forcing ω_{jk} equal to zero. In other words, we individually regress the components of C on X. The variable for hospital length of stay, which is continuous, it dichotomized at the sample median. The results, in the form of point and 95% interval estimates of $\theta_1,\ldots, \theta_{21}$ are plotted in the first column of Figure 2 under the heading "Crude analysis". For example, for C_1 which indicates female sex, we estimate θ_1 as 0.09 with 95% confidence interval (-0.01, 0.18), and this result is plotted accordingly. Figure 2 reveals that many covariates are associated with of treatment. In particular, treated patients are more likely to be treated with other heart failure medications.

Next we examine the performance of $\hat{Z}^{(PSA)}$ as a tool to reduce confounding. We fit the 21 regressions in equation (3) by substituting $\hat{Z} = \hat{Z}^{(PSA)}$ and compute point and interval estimates for $\theta_1,\ldots, \theta_{21}$. The results are given in the second column of Figure 2 under the heading "PSA". Here we see the that the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_{21}$ are close to zero because the same data are used to estimate the propensity scores and to check balance with respect to X. This illustrates the ability of PSA to control confounding. Compared to the crude analysis, Figure 2 reveals that stratifying on $\hat{Z}^{(PSA)}$ breaks the association between C and X, and thus reduces confounding.

Finally, we repeat the above regressions substituting $\hat{Z}^{(BPSA)}$ in place for $\hat{Z}$ in equation (3). We calculate point and interval estimates for $\theta_1,\ldots, \theta_{21}$ and plot the results in the final column of Figure 2 under the heading "BPSA". Figure 2 shows that stratifying on $\hat{Z}^{(BPSA)}$ does not yield the same degree of balance compared to stratifying on $\hat{Z}^{(PSA)}$. The point estimates $\hat{\theta}_1, \ldots, \hat{\theta}_{21}$ are generally closer to zero compared to the crude analysis, indicating that some of the association between X and C has been reduced. However, BPSA does not succeed in balancing

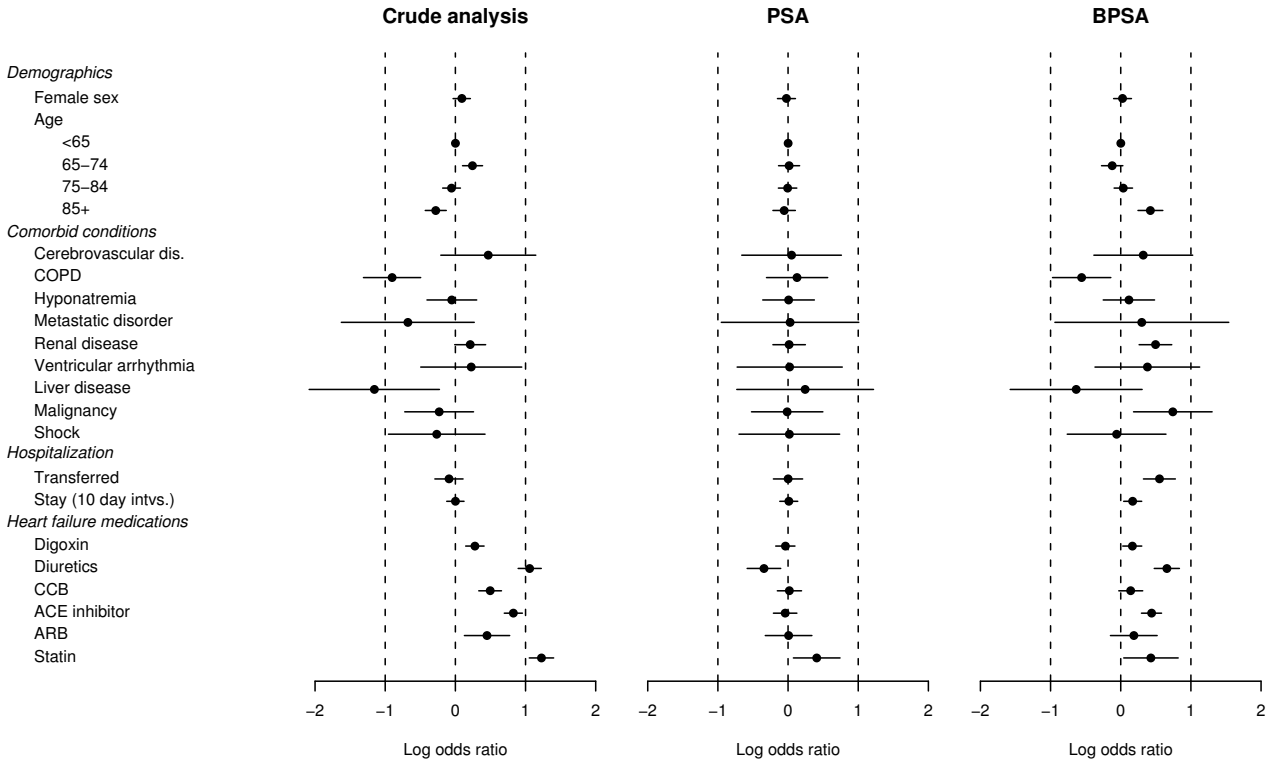

Figure 2

Balance with respect to treatment. Each row corresponds to the log odds ratio (95% CI) for the association between a covariate and treatment in either an unadjusted analysis, or after having adjusted for $\hat{Z}^{(PSA)}$ or $\hat{Z}^{(BPSA)}$.

the covariates as effectively as PSA. Therefore, BPSA appears to be less effective for controlling confounding than PSA.

The difficulty with this investigation is that it ignores associations between the co-variates and outcome variable. Figure 2 shows that adjusting for $\hat{Z}^{(PSA)}$ breaks the association between C_1,..., C_{21} and X. But it does not reveal if these variables are all equally important mortality risk factors. Recall that a covariate C_k is defined as a confounding variable if 1) $X \not\perp\!\!\!\perp C_k$, 2) $Y \not\perp\!\!\!\perp C_k|X$, and further that 3) C_k is not affected by X or Y [16]. It is useful to consider the relationship between the covariates and mortality if we wish to identify confounding bias. If certain components of C are more strongly associated with Y than others, then imbalances in Figure 2 may be misleading.

All of the covariates in the heart failure dataset are a priori known mortality risk factors [17]. But at issue is whether or not they are associated with mortality conditional on the estimated propensity score. The purpose of propensity techniques is to stratify the population into coarse subgroups within which treatment effect estimates are unconfounded. To get a clear picture of the performance of $\hat{Z}^{(BPSA)}$ and $\hat{Z}^{(PSA)}$ as tools to control confounding, we should explore the associations between C and Y conditional on the estimated propensity scores.

Balance with respect to the outcome

Review: Prognostic scores for control of confounding

Suppose that $\tilde{Z} = \tilde{Z}(C)$ denotes a scalar function of C. We say that there is *balance with respect to the outcome* if $Y \perp\!\!\!\perp C | \tilde{Z}, X = 0$. This conditional independence assumption says that, among untreated subjects with $X = 0$, the distribution of the outcome is determined by $\tilde{Z}$ and does not depend on C. Conceptually, the quantity $\tilde{Z}$ can be interpreted as a *prognostic score* in the sense that it is a scalar summary of the contribution of C to the outcome risk [18].

Hansen [18] recently introduced the notion of prognostic scores for control of confounding in observational studies. For the heart failure data, let Y_1 and Y_0 denote potential outcomes for death for a patient in the study. The quantity Y_1 models mortality at the end of follow up for treated patients and takes value one if the patient dies and zero otherwise. The quantity Y_0 models the corresponding potential outcome for death assuming the patient is untreated. Let $Y = Y_X$ denote the observed potential outcome.

A prognostic score $\tilde{Z}$ is defined as any scalar function of C with the property that

$$Y_0 \perp C \mid \tilde{Z}. \tag{4}$$

This equation says that $\tilde{Z}$ determines the distribution of the outcome among untreated subjects. See Hansen [18] for details.

Prognostic scores are analogous to propensity scores, and they have close connections to disease risk scores reviewed by Rosenbaum and Rubin [1], and Stürmer et al. [19]. Stratifying on a prognostic score removes confounding because it breaks the association between C and the outcome. Hansen [18] proves that when a prognostic score $\tilde{Z}$ is known, then this implies that we can control confounding by including it as a covariate in a regression model for the outcome. Effect measures calculated from $P(Y \mid X, \tilde{Z})$ have a causal interpretation.

The idea of prognostic scores give a theoretical basis for checking balance with respect to the outcome in the heart failure data. To compare the performance of $\hat{Z}^{(BPSA)}$ and $\hat{Z}^{(PSA)}$ for control of confounding, we can instead verify there is *balance with respect to the outcome*, meaning that $Y \perp C \mid \hat{Z}, X = 0$. If we see that there is balance with respect to the outcome, what this means is that the estimated propensity score $\hat{Z}$ breaks the association between the covariates and the outcome. Thus by stratifying on $\hat{Z}$, the covariates C cease to be outcome risk factors and are therefore no longer confounders.

A crucial feature of prognostic scores is that they do not require that $X \perp C \mid \tilde{Z}$. If $X \not\perp C \mid \tilde{Z}$, then effect measures computed from $P(Y|X, \tilde{Z})$ will nonetheless have a causal interpretation provided that equation (4) holds. In other words, just because a covariate summary score does not yield balance with respect to treatment does not imply that it cannot be used to control confounding. We may instead have $Y \perp C \mid \tilde{Z}, X = 0$ in which case $\tilde{Z}$ breaks the association between the confounders and outcome.

Balance with respect to mortality in the heart failure data

To study the ability of $\hat{Z}^{(BPSA)}$ and $\hat{Z}^{(PSA)}$ to induce balance with respect to the outcome in the heart failure data, we employ a similar empirical investigation strategy to the one described above. We fit models of the form

$$\text{logit}[P(Y = 1 \mid C_k, X, Z)] = \phi_k + \sum_{j=1}^{l=3} \omega_{jk} g_j\{Z\} + \rho_k C_k \quad \text{for } k = 1, \ldots, 21 \tag{5}$$

where C_k denotes the k^{th} component of C, the parameter ϕ_k is a regression intercept, and $\hat{Z}$ is a propensity score estimate. Equation (5) is identical to the outcome regression model of equation (1) with $X = 0$, except that we now include the additional covariate C_k in the model. It assesses whether or not we should include the covariates C in the model in addition to the estimated propensity score. If the estimated propensity score $\hat{Z}$ induces balance with respect to the outcome, meaning that $Y \perp C_k \mid \hat{Z}, X = 0$, then we should have $\rho_1 = \rho_2 = \ldots = \rho_{21} = 0$. We may once again compute point estimates $\hat{\rho}_1, \ldots, \hat{\rho}_{21}$ and study the extent to which they depart from zero.

First we calculate the crude associations between Y and $C_1, \ldots, C_{21}$ among untreated subjects. This is accomplished by fitting the $p = 21$ regressions in equation (5) while forcing $\omega_{jk} = 0$. In order words, we regress Y on each of the individual components of C. The results in the form of point and 95% interval estimates of $\rho_1, \ldots, \rho_{21}$ are given in the first column of Figure 3 under the heading "Crude analysis". We see that many covariates are strong risk factors for mortality. For example, most of the comorbid conditions are associated with increased risk of death.

Next we examine the performance of $\hat{Z}^{(BPSA)}$ and $\hat{Z}^{(PSA)}$ as tools to reduce confounding. We fit the 21 regressions in equation (5) substituting either $\hat{Z}^{(BPSA)}$ or $\hat{Z}^{(PSA)}$ in place of $\hat{Z}$, and computing the corresponding inferences for $\rho_1, \ldots, \rho_{21}$. The results are given in the second and third columns of Figure 3 under the headings "PSA" and "BPSA" respectively.

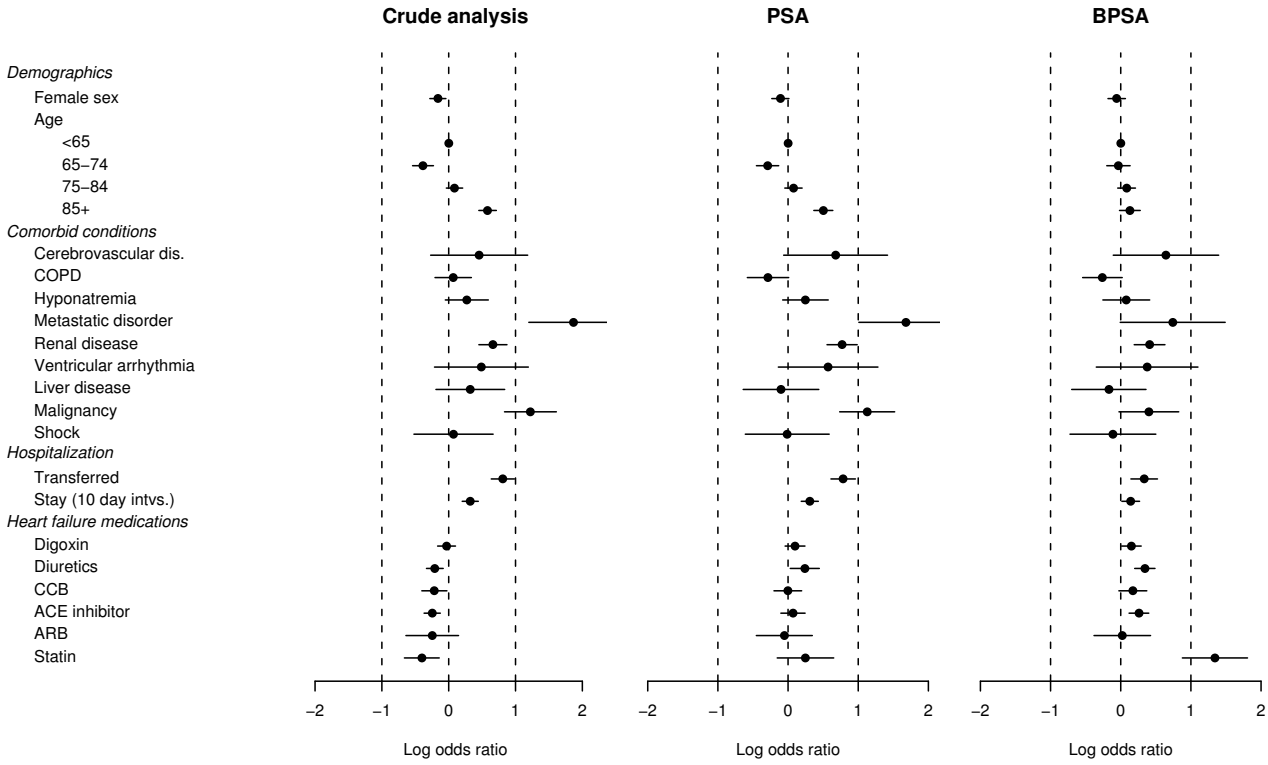

Figure 3

Balance with respect to the outcome. Each row corresponds to the log odds ratio (95% CI) for the association between a covariate and mortality, within treatment groups, in either an unadjusted analysis, or after having adjusted for $\hat{Z}^{(PSA)}$ or $\hat{Z}^{(BPSA)}$.

Figure 3 indicates that BPSA produces greater balance with respect to mortality compared to PSA. The point estimates of $\rho_1,\ldots,\rho_{21}$ are shifted towards zero. As it stands, the BPSA model *assumes* that the propensity scores achieve balance with respect to the outcome, and the rightmost column of Figure 3 has diagnostic value in supporting or refuting this assumption. So for the present data we see that the assumption is not bad, but not perfect. For example, consider the variable metastatic disorder. In Figure 3, under the heading "Crude analysis" we see that this variable is the strongest predictor of mortality in the heart failure dataset with an estimated log odds ratio of greater than 2. Stratifying on $\hat{Z}^{(BPSA)}$ breaks much of this association, while stratifying on $\hat{Z}^{(PSA)}$ does not.

Table 2: Summary statistics for the distribution of log odds ratios depicted in Figure 2 and Figure 3.

	Log odds ratios			
	Median	**IQR**	**Mean**	**Variance**
Balance with respect to treatment				
Crude	0.09	0.68	0.09	0.34
PSA	0.02	0.04	0.02	0.02
BPSA	0.20	0.39	0.19	0.12
Balance with respect to outcome				
Crude	0.08	0.70	0.25	0.31
PSA	0.24	0.58	0.31	0.24
BPSA	0.15	0.35	0.23	0.12

To give a clearer comparison of the tradeoffs between BPSA and PSA, Table 2 gives summary statistics for the distribution of the log odds ratios from Figures 2 and 3. For PSA, we see the method gives good balance with respect to treatment because the $\hat{\rho}_j$ are all close to zero. By comparision, BPSA does a better job of reducing the magnitude of the associations between the confounders and outcome that are depicted in Figure 3. For BPSA the sample mean and median are lower compared to PSA. The associations between C and Y are weaker overall after adjusting for $\hat{Z}^{(BPSA)}$. Similarly, the variance $\sum_{j=1}^{21} \frac{(\rho_j - \bar{\rho})^2}{21-1}$ is 0.12 for BPSA versus 0.24 for PSA. Table 2 does not point decisively towards the superiority of either method. Instead it describes the merits and tradeoffs of the Bayesian approach of using the outcome variable to estimate the propensity scores.

Conclusion

In the population of heart failure patients, confounding from C is driven by associations between C and Y, as well as by associations between C and X. By conditioning on $\hat{Z}^{(BPSA)}$, the Bayesian propensity score method yields strata where we have roughly $Y \perp C|X, Z$. Confounding is reduced because C are no longer strong mortality risk factors. By fitting regressions for X and Y simultaneously, BPSA treats the propensity score as a predictor for the outcome. The tradeoff is that BPSA is less successful in balancing the confounders with respect to the treatment variable.

In contrast, PSA estimates propensity scores from the marginal model for X given C. The method handles all components of C similarly, regardless of the strength of their association with Y. Figure 2 reveals that PSA balances all covariates equally well, but the approach may be overly pessimistic. Some covariates are more important mortality risk factors than others. It should be emphasized that neither method is able to reduce confounding from unobserved covariates.

A limitation of BPSA is that the propensity score estimates cannot be used to study multiple outcomes. Whereas traditional propensity scores ignore the outcome variable, Bayesian propensity scores are outcome specific. Equation (1) makes use of modelling assumptions for the relationship between Z and the particular outcome Y under investigation. Indeed the strategy of fitting models for the treatment and outcome simultaneously goes against the idea of setting up the study design and analysis without access to the outcome [20]. Nonetheless, some authors argue that the performance of propensity score techniques is improved by making use of the outcome data. If our objective is to control confounding, then variables that are weakly associated with the outcome are less important in propensity score modelling, regardless of whether they are strong predictors of treatment [6,12,20,21].

From a substantive point of view, one can argue that equation (1) is not realistic in the sense that we do not expect to have $Y \perp C|X, Z$. Nonetheless, in regression adjustment for the propensity score, the investigator must choose a model for the relationship between the outcome and the propensity score. Popular choices include stratifying on subclasses of the propensity score or assuming a linear relationship. The model should be based on genuine beliefs about the relationship between the propensity score and the outcome. Further discussion of model based regression adjustment for the propensity score is given by Rosenbaum and Rubin [1].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LM conducted the analysis, interpretation of results, and drafting and revising the manuscript. PG and PA contribute to conceiving and interpreting the analysis as well as drafting and revising the manuscript. AL contributed to acquisition of the data and revising the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Code for fitting Bayesian propensity analysis to a toy synthetic dataset. *Computer code for implementing BPSA in the software package R.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-5573-6-5-S1.R]

References

1. Rosenbaum PR, Rubin DB: **The central role of the propensity score in observational studies for causal effects.** *Biometrika* 1983, **70:**41-57.
2. Austin PC, Mamdani MM: **A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use.** *Statistics in Medicine* 2005, **25:**2084-106.
3. McCandless LC, Gustafson P, Austin PC: **Bayesian propensity score analysis for observational data.** *Statistics in Medicine* 2009, **28:**94-112.
4. Hill JL, McCulloch RE: **Bayesian nonparametric modeling for causal inference.** *Journal of the American Statistical Association* 2009 in press.
5. Hoshino T: **A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm.** *Computational Statistics and Data Analysis* 2008, **52:**1413-29.
6. Austin PC, Grootendorst P, Anderson GM: **A comparison of the ability of different propensity score models to balance meas-**

ured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine* 2007, **26:**734-53.
7. Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV: **The use of the propensity score for estimating treatment effects: Administrative versus clinical data.** *Statistics in Medicine* 2005, **24:**1563-78.
8. McCandless LC, Gustafson P, Levy AR: **Bayesian sensitivity analysis for un-measured confounding in observational studies.** *Statistics in Medicine* 2007, **26:**2331-47.
9. Foody JAM, Farrell MH, Krumholz HM: **β-blocker therapy in heart failure: Scientific Review.** *Journal of the American Medical Association* 2002, **278:**883-9.
10. Glynn RJ, Knight EL, Levin R, Avorn J: **Paradoxical relations of drug treatment with mortality in older persons.** *Epidemiology* 2001, **12:**682-9.
11. McCandless LC, Gustafson P, Levy AR: **A sensitivity analysis using information about measured confounders yielded improved assessments of uncertainty from unmeasured confounding.** *Journal of Clinical Epidemiology* 2008, **61:**247-55.
12. Rubin DB, Thomas N: **Matching using estimated propensity scores: Relating theory to practice.** *Biometrics* 1996, **52:**249-64.
13. R Development Core Team: *R: A language and environment for statistical computing* 2004 [http://www.R-project.org]. R Foundation for Statistical Computing: Vienna ISBN 3-900051-00-3.
14. Gelman A, Carlin JB, Stern HS, Rubin DB: *Bayesian Data Analysis* 2nd edition. Chapman Hall/CRC:New York; 2003.
15. Imai K, van Dyk DA: **Causal inference with general treatment regimes: Generalizing the propensity score.** *Journal of the American Statistical Association* 2004, **99:**854-66.
16. Hernbán MA, Hernández-Diaz S, Werler MM, Mitchell AA: **Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology.** *American Journal of Epidemiology* 2002, **155:**176-84.
17. Polanczyk CA, Rohde LE, Philbin EA, Di Salvo TG: **A new casemix adjustment index for hospital mortality among patients with congestive heart failure.** *Medical Care* 1998, **36:**1489-99.
18. Hansen BB: **The prognostic analogue of the propensity score.** *Biometrika* 2008, **95:**481-88.
19. Stürmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ: **Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: Nonsteroidal anti-inflammatory drugs and short-term mortality in the elderly.** *American Journal of Epidemiology* 2005, **161:**891-9.
20. Rubin DB: **For objective causal inference, design trumps analysis.** *Annals of Applied Statistics* 2008, **2:**808-40.
21. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stümer T: **Variable selection for propensity score models.** *American Journal of Epidemiology* 2006, **163:**1149-56.